An Overview of the Sandia Genomes to Life Project

"Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling"

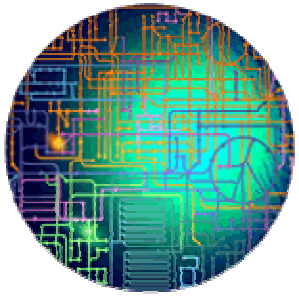
Grant S. Heffelfinger, PhD

Principle Investigator

Sandia National Laboratories

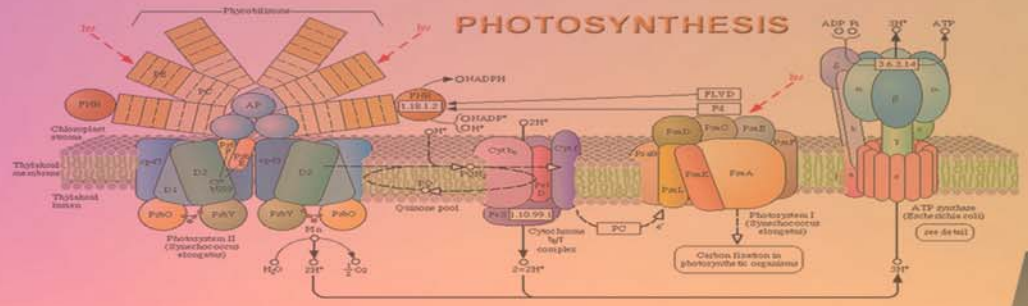
Albuquerque, New Mexico 87185

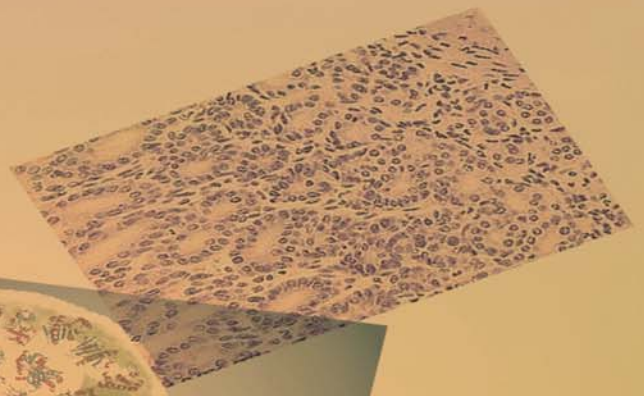
gsheffe@sandia.gov



Sandia National Laboratories

Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling





Sandia National Laboratories
Oak Ridge National Laboratory
Lawrence Berkley National Laboratory
Los Alamos National Laboratory
U Michigan
UC Santa Barbara
U Illinois Urbana/Champaign
The National Center for Genome Resources
Scripps Institution of Oceanography
The Molecular Science Institute
Joint Institute for Computational Science

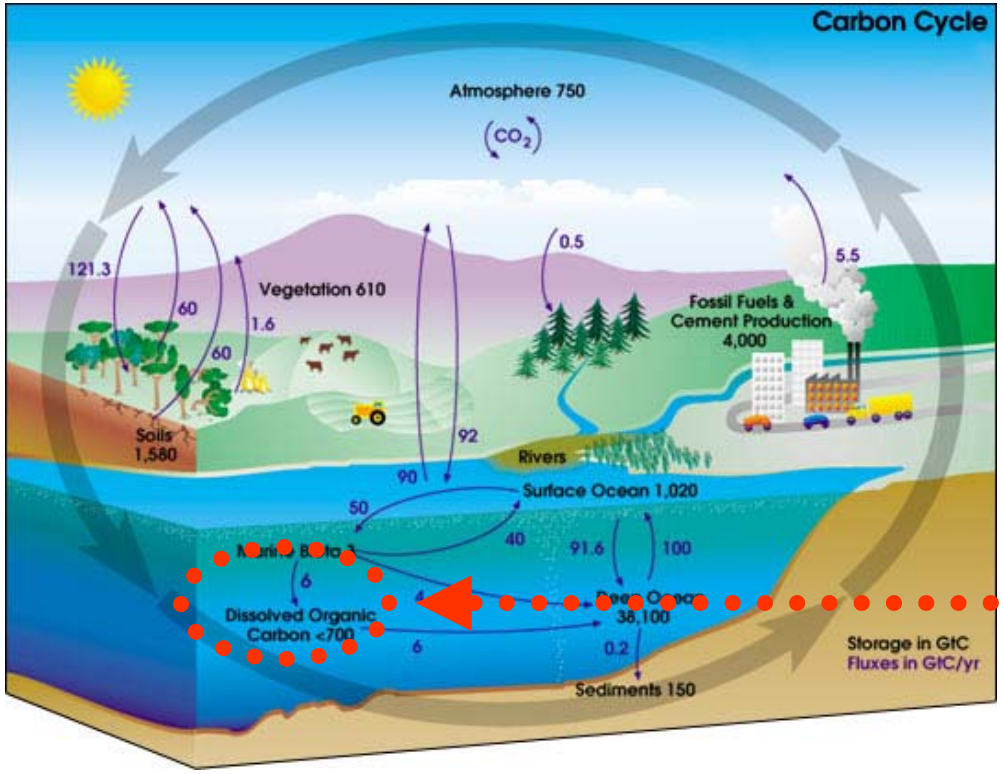
Genomes to Life

Program Goals

1. Identify and characterize the molecular machines of life – the multiprotein complexes that execute cellular functions and govern cell form
2. Characterize gene regulatory networks
3. Characterize the functional repertoire of complex microbial communities in their natural environments at the molecular level
4. Develop the computational methods and capabilities to advance understanding of complex biological systems and predict their behavior

Sandia's Genomes to Life Project

"Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling"



The major **goal** of this effort is to develop *computational methods and capabilities* to advance understanding of complex biological systems and predict their behavior.

The initial target for the **development and testing** of the new methods and tools is *Synechococcus* Sp., an ocean bacteria which plays a central role in climate change by fixing atmospheric carbon.

The major **biological objective** of this work is to *elucidate the relationship of the Synechococcus genome* to *Synechococcus'* relevance *to global carbon fixation*.

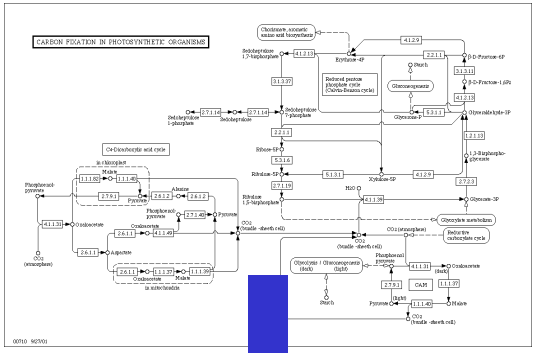
Sandia National Laboratories

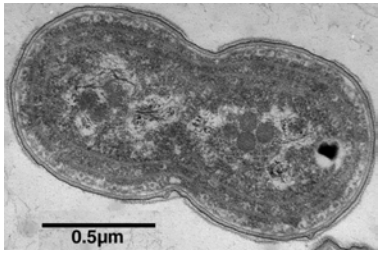
Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Carbon Fixation in *Synechococcus*

A Computational Decomposition of the Problem

- Identify candidate proteins involved in carbon fixation through gene expression data analysis, regulatory binding site prediction, and operon/regulon structure prediction
- Identify protein interactions through analysis of affinity data and public protein-protein interaction data
- Protein structure prediction through Rosetta-type algorithms and refinements
- Ellucidate gene regulatory pathways via systematic inference methods
- Link to cellular and macroscopic response
- Experimental verification
- *Model refinement through an iterative process of computation and experiments*



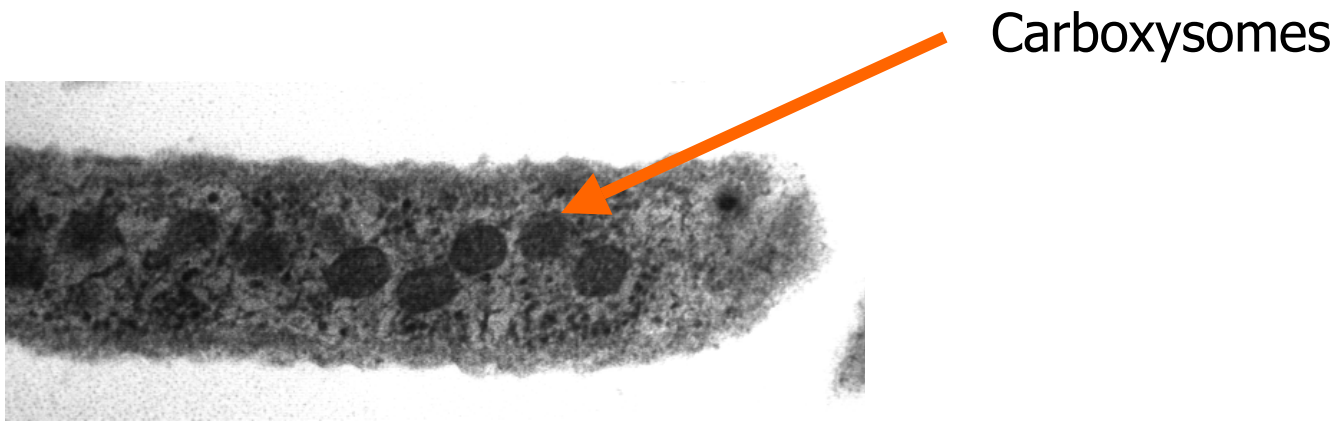


Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Carbon Fixation & Molecular Machines

Carboxysome, ABC Transporters, and Histidine Kinase-Response Regulators



Carboxysomes have been experimentally characterized

- at least ten polypeptides present
- two inside the core (structures known)
- > 6 or 7 are in the shell (structures not known)

Our computational and experimental efforts will focus on molecular machines key to the carbon fixation process in *Synechococcus*.

Background & Significance of Goal 4

- Biology is undergoing a major transformation that will be enabled and ultimately driven by computation.
- High-performance computing is essential to the high-throughput experimental approach to biology that has emerged in the last 10 years.
- Ease of use and coupling between geographically and organizationally distributed people, data, software, and hardware is critical.
- Work environments must be conceptually integrated "knowledge enabling" environments that couple diverse sets of distributed data, advanced informatics methods, experiments, and modeling and simulation.

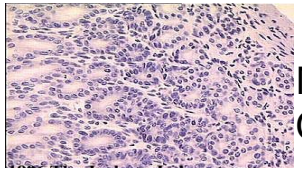
Goal 4 Spans the Genomes to Life Program

Microbial Complexity

Ecological Processes and Populations

Evolutionary processes

Simple ecosystems

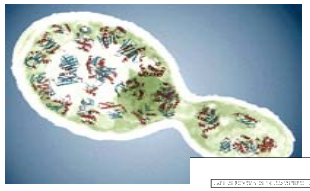


Multi-organism Community behavior

Cellular Communities

Single microbe community

Cellular Processes

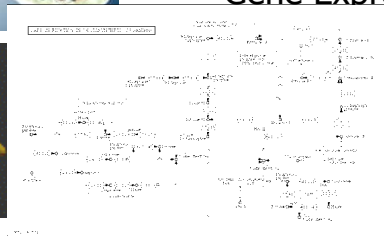


Whole Cell Modeling

Catalog complexes

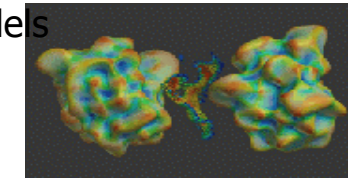
Gene Expression Networks

Regulatory Pathways



Gene Regulation Pathways

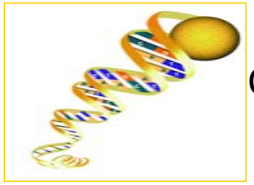
Steady state metabolic models



Molecular Machines

Comparative Protein Analysis

Protein Complexes



Protein Sequence Prediction

Genome Comparisons

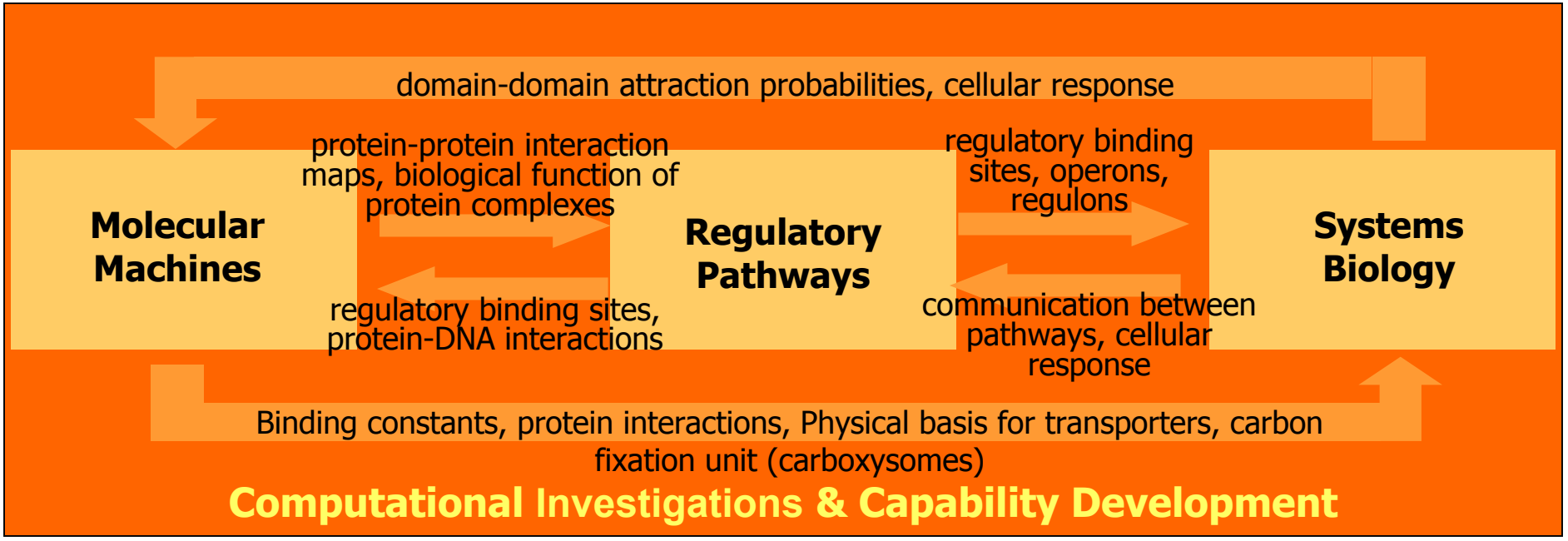
Protein Structure Modeling

1 10 100 1000 Tflops

Computing and Information Requirements

Sandia National Laboratories

Three Synergistic Computational Biology Efforts Form the Core of This Effort



Two Additional Efforts Support the Computational Biology Core

phage display data

gene expression data

protein network

protein networks

regulatory network characterization

Matlab-like systems biology tool

rankings of ligand binding affinities

ligand binding data

Gene regulation networks

Inferred networks

microarray data analysis

Experimental Investigation & Data Generation

data mgt/clustering for gene expression, protein interactions, & regulatory networks

Work Environments & Computational Infrastructure

Biology web portal/Genomic Integrated Supercomputing Toolkit

Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Molecular Machines

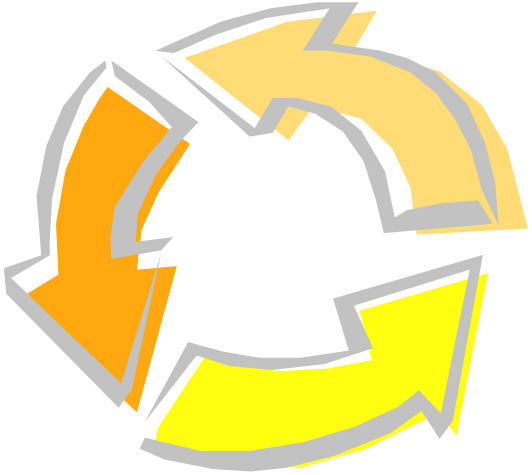
First Element of the Computational Core

Molecular Machines

Regulatory Pathways

Systems Biology

Computational Investigations & Capability Development



Experimental Investigation & Data Generation

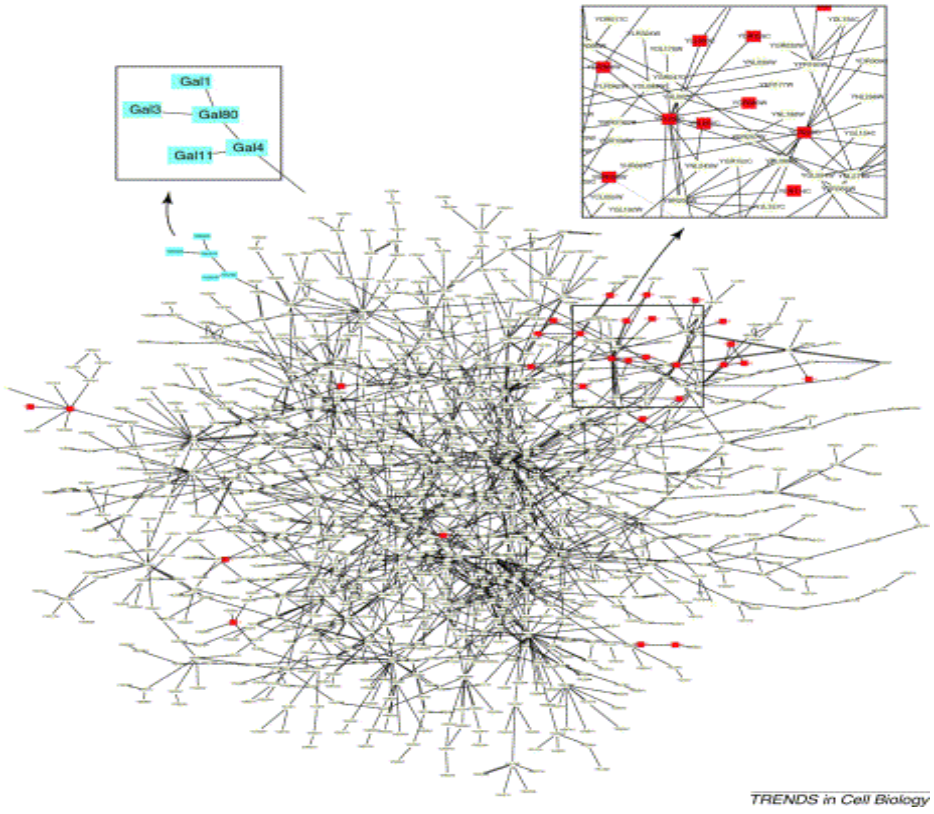
Work Environments & Computational Infrastructure

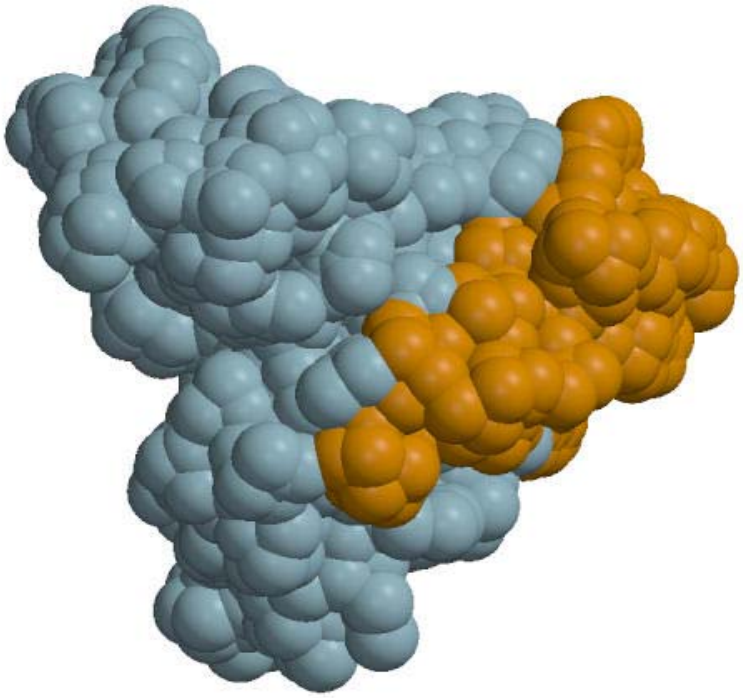
Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Molecular Machines

Discovery and Validation of Protein-protein Complexes





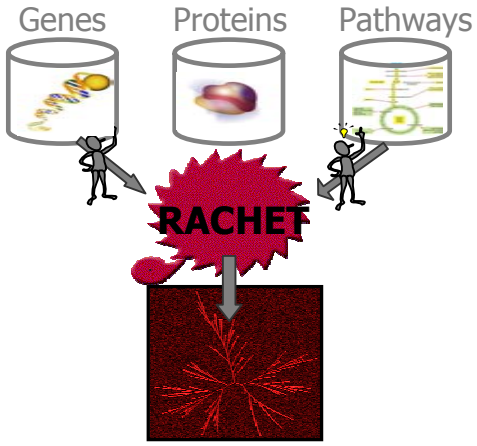
Bioinformatics & Data Mining	Docking with Rosetta Scoring	Molecular Biophysics

Sandia National Laboratories

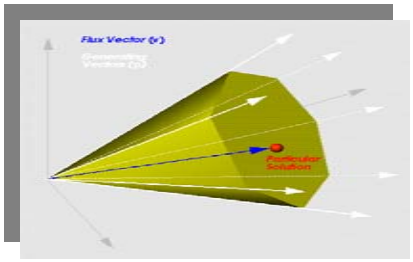
Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Bioinformatics & Data Mining

Discovering of Protein-protein Interactions With "Knowledge Fusion"



Clustering algorithms for distributed databases

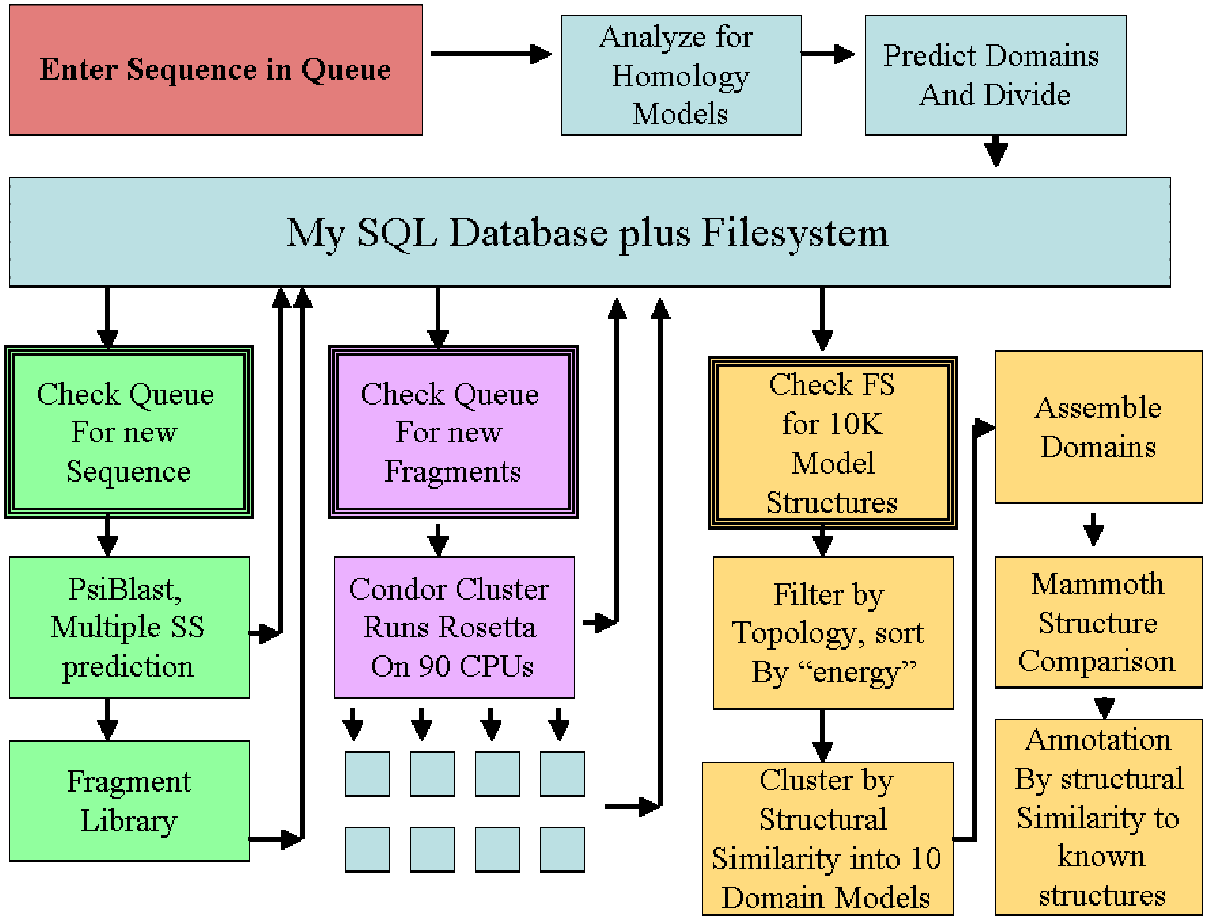


- An order of magnitude larger systems
- Memory: reduced by over 90%
- Time: reduced from days to hours

- Develop categorical analysis tool combing several genome context data sources for analysis of protein-protein interaction. Create catalog of proteins in *Synechococcus* that are relevant to specific metabolic pathways.
- Incorporate structural information in mining algorithms for protein-protein interactions (e.g. **P**rotein **I**nteraction **C**lassification by **U**nlikely **P**rofile **P**air (PICUPP)
- Extend algorithmic applicability to the scale of whole genomes through implementation of optimized versions suited for Terascale computers.

Protein-protein Docking With Rosetta-like Methods

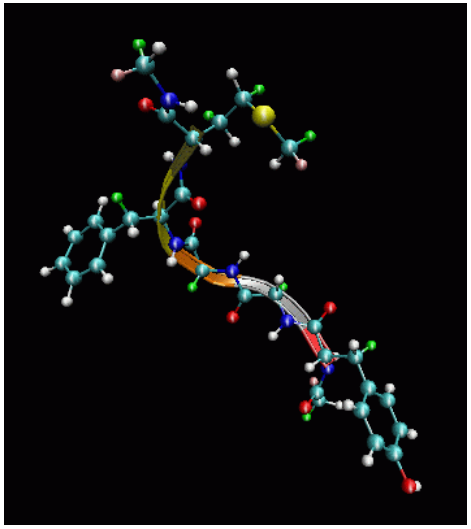
Robetta structure prediction pipeline.



- Extend Rosetta to protein-protein complexes & HPC environments
- Incorporate experimental constraints for de novo sequencing via mass spectrometry (e.g. "probability profiles" (Fridman et al., "Probability Profiles - Novel Approach in Mass Spectrometry de novo Sequencing," **Proceedings of the IEEE Computer Society Bioinformatics Conference, Stanford, August 2003**.)
- New sampling techniques and molecular biophysics for increased discrimination of native folds.

Molecular Biophysics Approaches

New Algorithms, Simulation Methods, & Massively Parallel Computing Essential



- Perform large-scale MD, parallel tempering, and docking of phage display ligand/protein complexes. (implementation nearly working, post-tempering tools under development - 3 target calculations planned: 1) ligand conformation to match PDB and docking results, 2) ab initio prediction of phage display, and 3) ligand conformations relaxation/comparison of Rosetta conformations)
- Interface with Rosetta to narrow conformational search. Prototype the whole pipeline on important protein-protein complexes in *Synechococcus*.
- Extend classical DFT methods, and supporting parallel algorithms and solvers, for modeling functionality of membrane transporters: solution properties, charge polarizations, ion properties.

Molecular Machines

Ultimate Goal

Integrated computational tools for the exploration of protein-protein interactions on a genomic scale

Bioinformatics & Data Mining

Molecular Biophysics

Docking with Rosetta Scoring

- Prototyping through investigation of the important protein-protein interactions in *Synechococcus*
- Advances in fundamental understanding of protein-protein interaction with wide application to other microbes

Regulatory Pathways

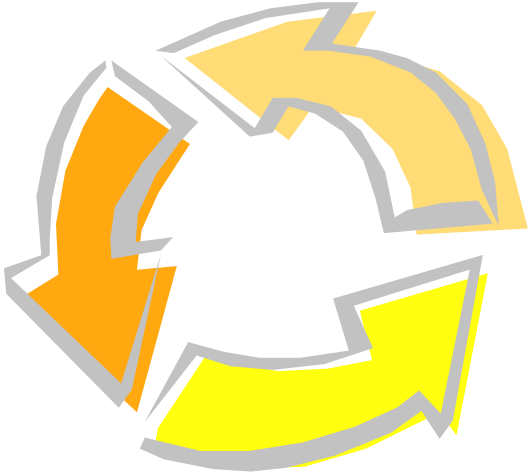
Second Element of the Computational Core

Molecular Machines

Regulatory Pathways

Systems Biology

Computational Investigations & Capability Development



Experimental Investigation & Data Generation

Work Environments & Computational Infrastructure

Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Regulatory Pathways

Develop Reliable & Systematic Methods to Infer Regulatory Pathways

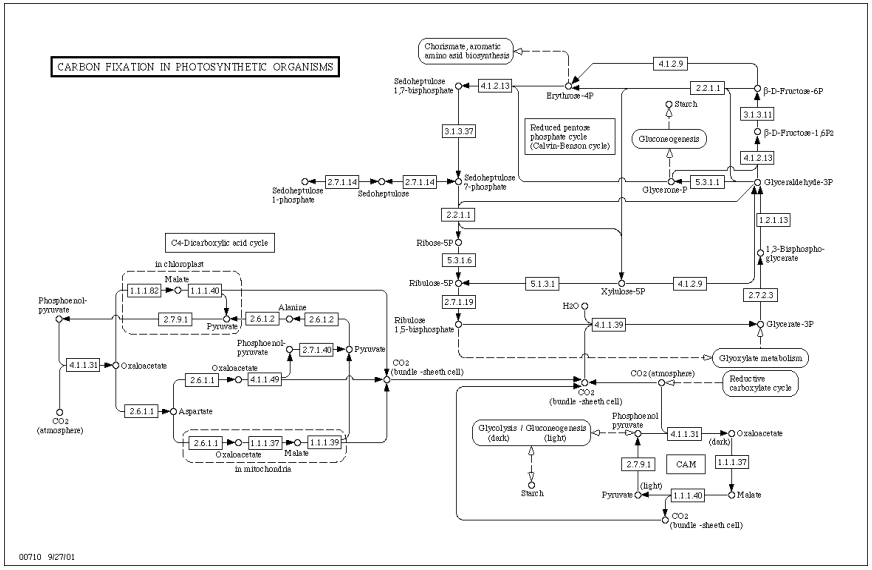
Regulatory networks are responsible for control of biological functions at both cellular and molecular levels

Information available for deciphering regulatory pathways

regulatory binding sites

operons/regulons

evolutionary data derivable from genomic sequences



two-hybrid data

partial pathways from other genomes

microarray gene expression data

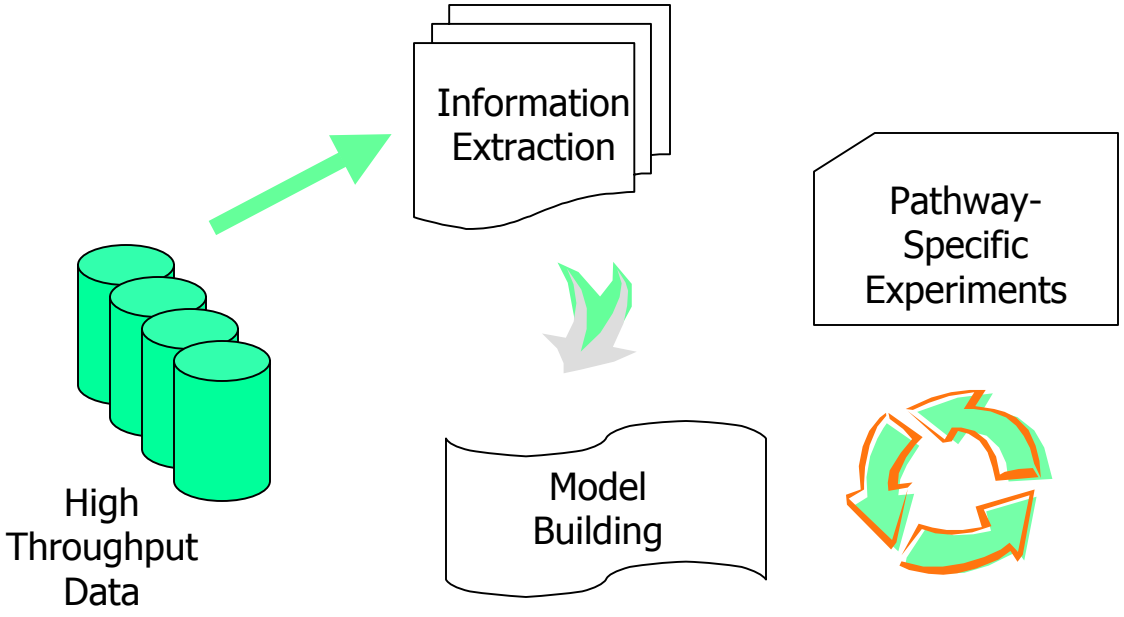
pathway-specific experimental data

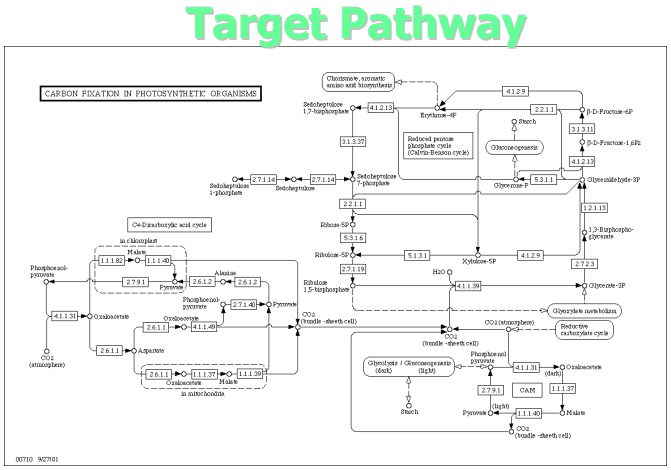
biological domain knowledge

Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

From Data to Pathways





- Data: use existing genomic sequences, gene expression data, protein-protein interaction data, partial pathways from other genomes, and on-line literature searches
- Carryout information extraction for target pathway
- Build pathway models consistent with derived information and biological knowledge (initially semi-automatic, later more automated)
- Design pathway-specific experiments to collect data to "fill the gaps" and validate pathway models

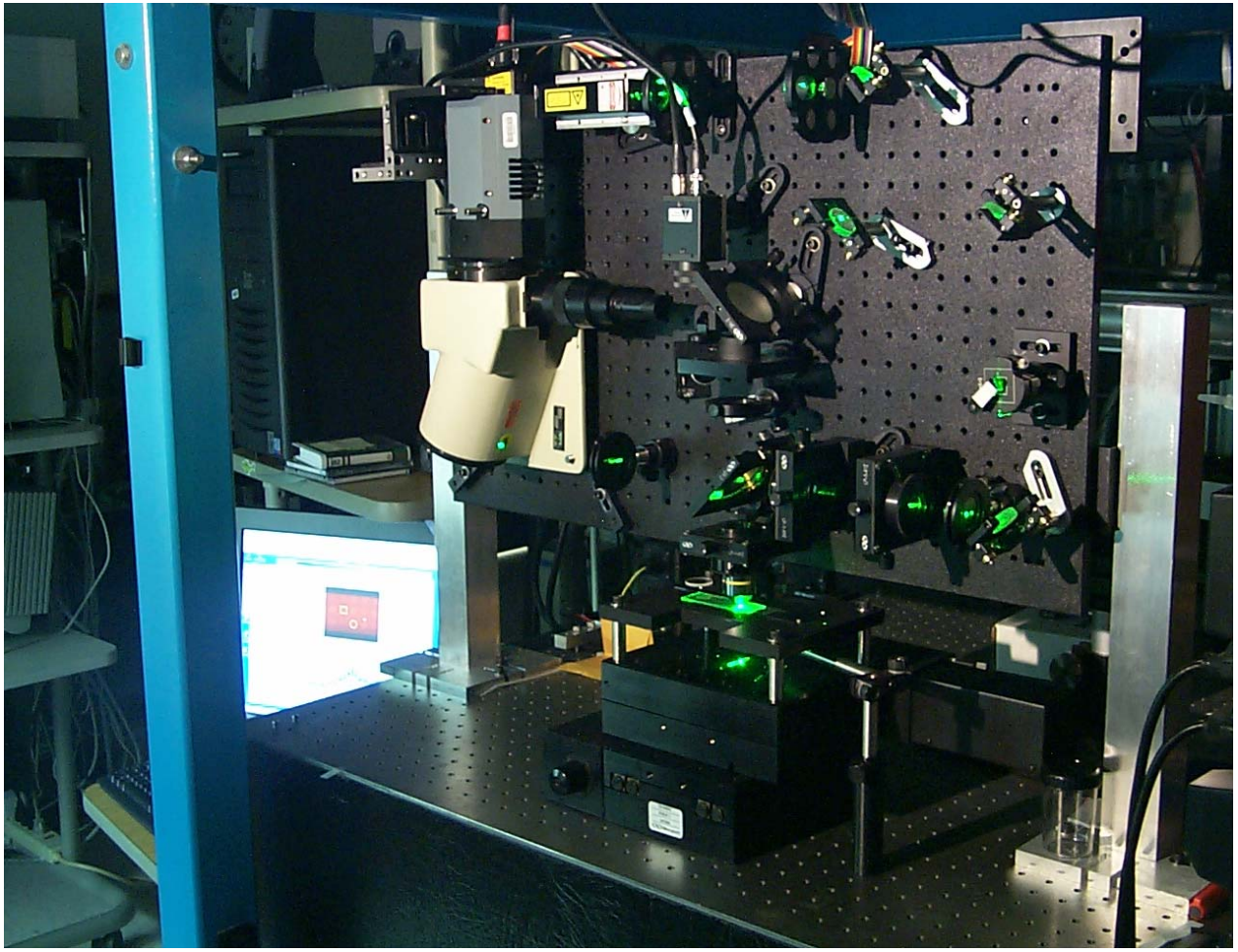
Regulatory Pathways

Progress

- Predicted a signaling/regulatory network for the phosphorus assimilation pathway in *Synechococcus* WH8102 through data mining and computational modeling. Work underway on two additional signaling/regulatory networks for nitrate and carbon assimilation in *Synechococcus* WH8102.
- Predicted protein-protein interaction map at genome scale for *Synechococcus* WH8102 via data mining and information fusion.
- Completed genome-scale protein structure/function predictions on all orfs of *Synechococcus* sp. and two related genomes *Prochlorococcus* MIT and MED (all prediction results are at http://compbio.ornl.gov/PROSPECT/syn/

Microarray Analysis

Hyperspectral Imaging

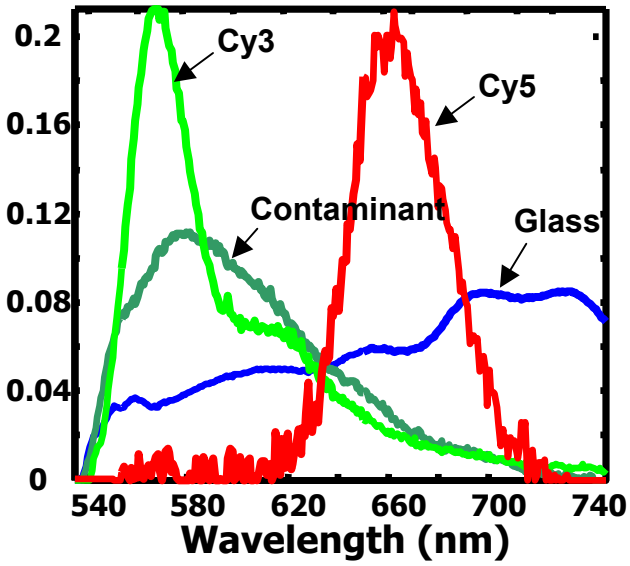


Collect an entire spectrum at each pixel, use multivariate data analysis to separate overlapped spectra into pure spectra of each emitting species and generate corresponding concentration maps

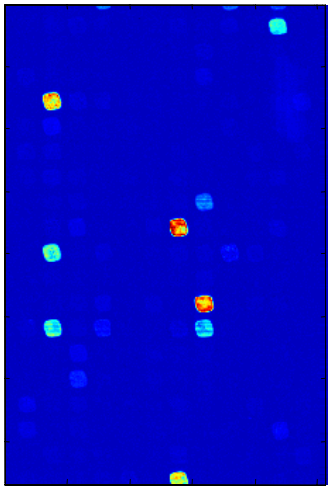
- Increased dynamic range
- Single laser, single scan, many fluorophores
- Increased reliability of microarray data

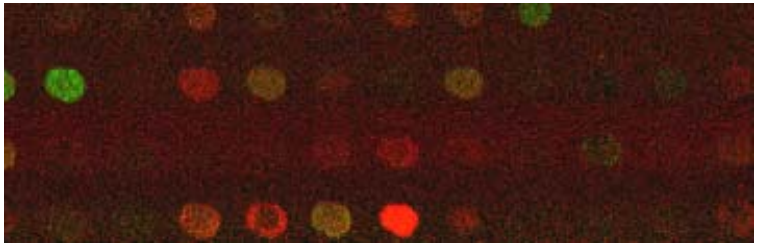
Microarray Analysis

Multivariate Curve Resolution



4096 X Wavelet Compression Projected To Full Resolution





- Reduce spectral & spatial resolution (>8x)
- Whole slide image <0.5GB
- Spectral & 2-D spatial compression (>2500x)
- MCR applied to fully compressed data w/o loss
- Project back to full spatial resolution of original data
- Data read + computation time <7 sec

Systems Biology

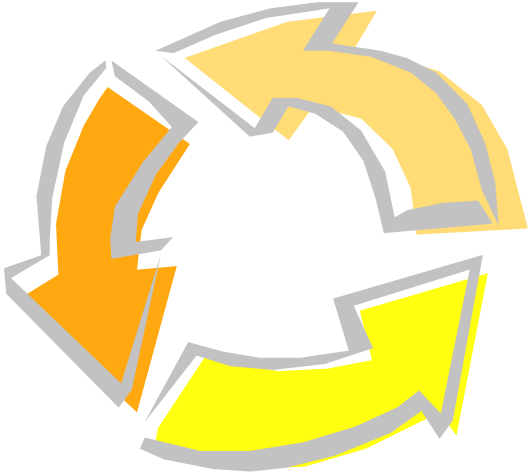
Third Element of the Computational Core

Molecular Machines

Regulatory Pathways

Systems Biology

Computational Investigations & Capability Development



Experimental Investigation & Data Generation

Work Environments & Computational Infrastructure

Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Systems Biology

New Simulation Methods, Algorithms, & Massively Parallel Computing Are Essential

Motivation

Linking the genome to a complex system response (e.g. cellular/environmental).

Goals

- Develop and prototype new methods for linking the genome, knowledge of molecular machines, and regulatory network understanding to build a more fundamental understanding of the microbe as a whole.
- Explore hierarchical bio-feedback approaches to modeling *Synechococcus* that will help build a fundamental understanding of the process of carbon-sequestration, from the genome to the environment.

Systems Biology

Research Strategy

Protein Interaction Network Inference and Analysis

Compute domain-domain attraction probabilities from phage display data, molecular simulation and protein interaction networks & use to sample (construct) a set of self-similar graphs which best optimizes these properties.

Spatial & Temporal Models of Biochemical Interactions

Develop simulation methods, companion algorithms and MP implementations for biochemical interactions via methods for evolving protein interaction networks forward in time via probabilistic, rule-based approach ("stochastic particle approach")

Couple network simulations to particle based simulations to capture the spatial aspects of the interactions.

Investigate continuum approaches for higher concentration ionic species, such as inorganic and organic carbon, for modeling reaction diffusion equations on realistic *Synechococcus* geometries.

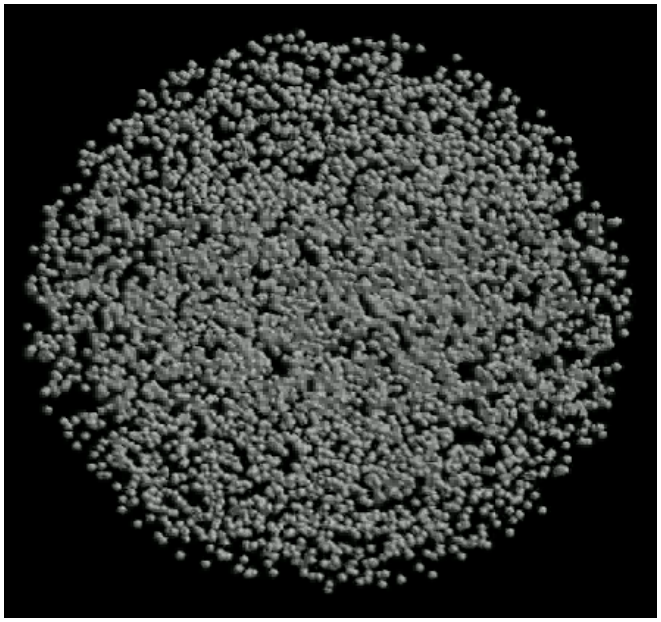
Hierarchical Models of the *Synechococcus* Carbon Sequestration Process

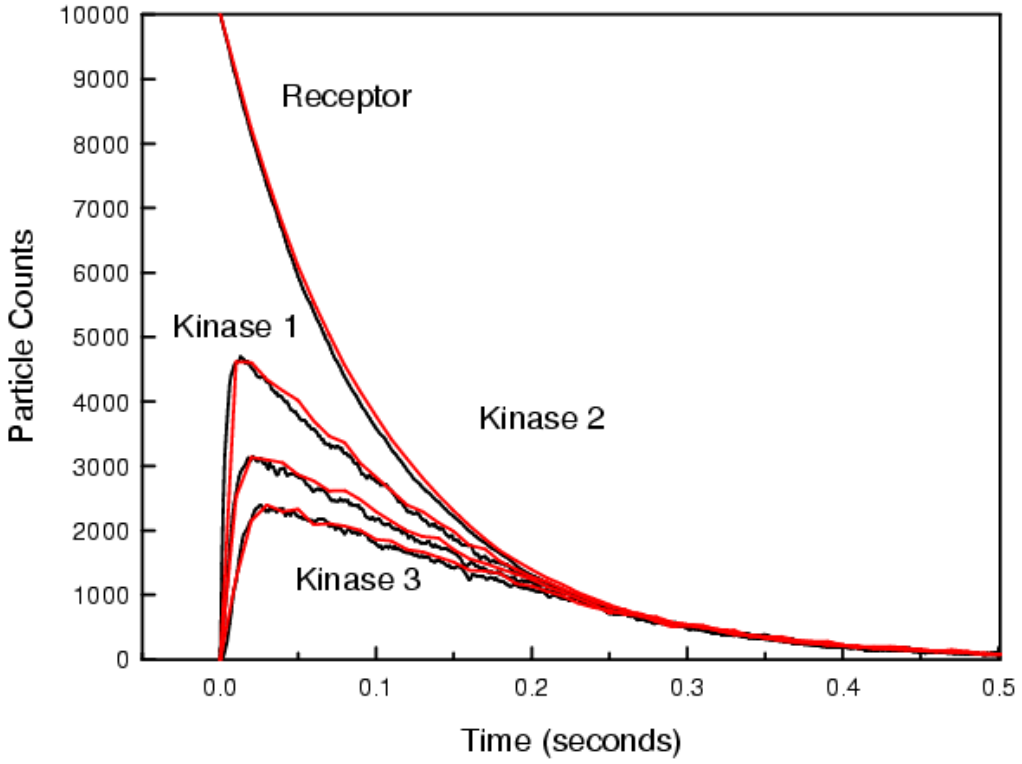
Construct the levels of the hierarchical model based on a biological understanding of *Synechococcus,* develop mathematical models coupling levels, evolve and refine the model to yield an understanding of how genetic and environmental factors affect the ability of *Synechococcus* to sequester inorganic carbon.

Stochastic Particle Dynamics

Model of Simple Signaling Cascade

- 9 species (1 receptor, 3 kinase, phosphatase), 7 reactions: 5000 particles, 10000 timesteps for 1 sec real time





Experiments & Data Generation

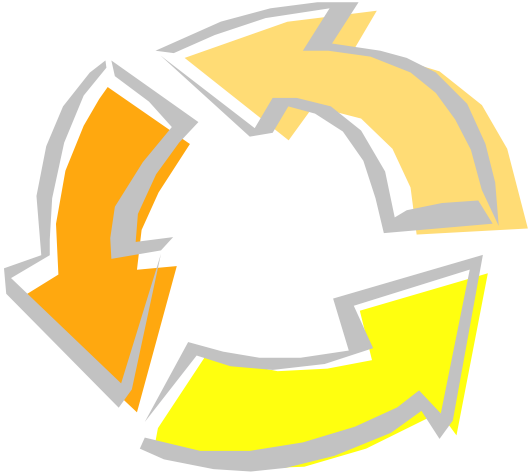
Complements the Computational Core

Molecular Machines

Regulatory Pathways

Systems Biology

Computational Investigations & Capability Development



Experimental Investigation & Data Generation

Work Environments & Computational Infrastructure

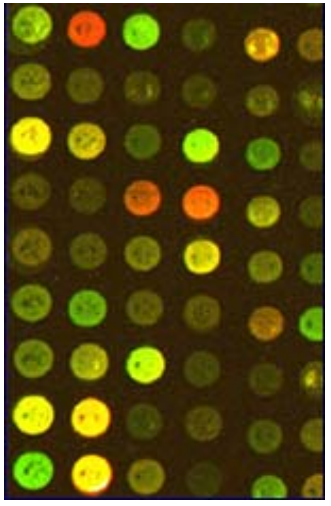
Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Experiments & Data Generation

Synechococcus Sp. WH8102

- Easily cultured under natural or artificial conditions.
- Amenable to biochemical and genetic manipulation.
- Established DNA microarrays.
- Comparative genomics can be carried out with *Prochlorococcus*
- Our effort includes marine biologist Brian Palenik, UCSD, collaborator in sequencing and annotating the *Synechococcus* genome.



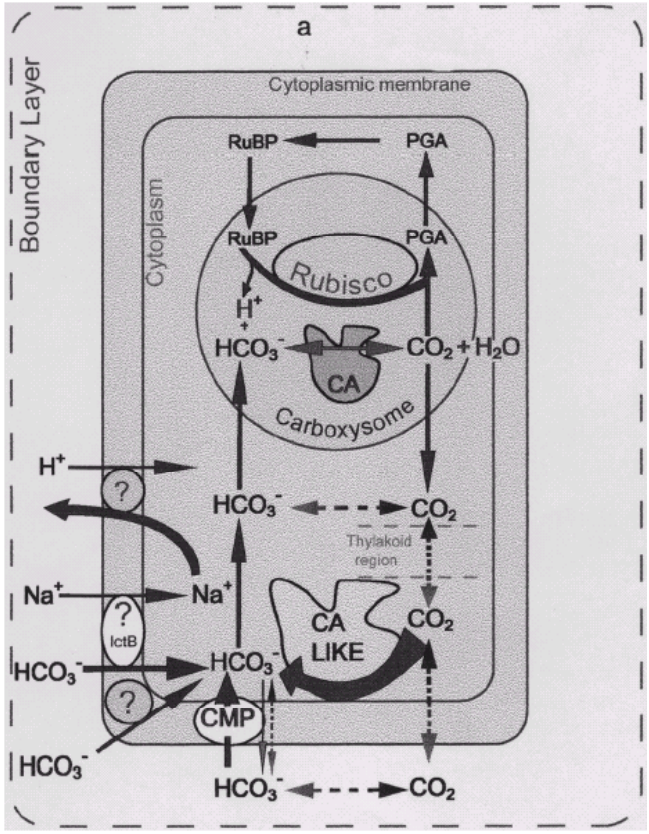
Experimental Goals

Elucidate Molecular Machines and Regulatory Networks

- Characterize complexes critical in carbon fixation.
- Characterize binding domains that mediate protein-protein interactions.
- Investigate co-regulation of complex genes & characterize protein expression levels.

Methods

- Affinity Purification/Mass Spectrometry
- Phage Display/ELISA/Yeast 2-hybrid
- Expression Library Screening
- Gene & Protein Microarrays
- New hyperspectral microarray scanner improves accuracy, dynamic range, and reliability of microarray experiments.

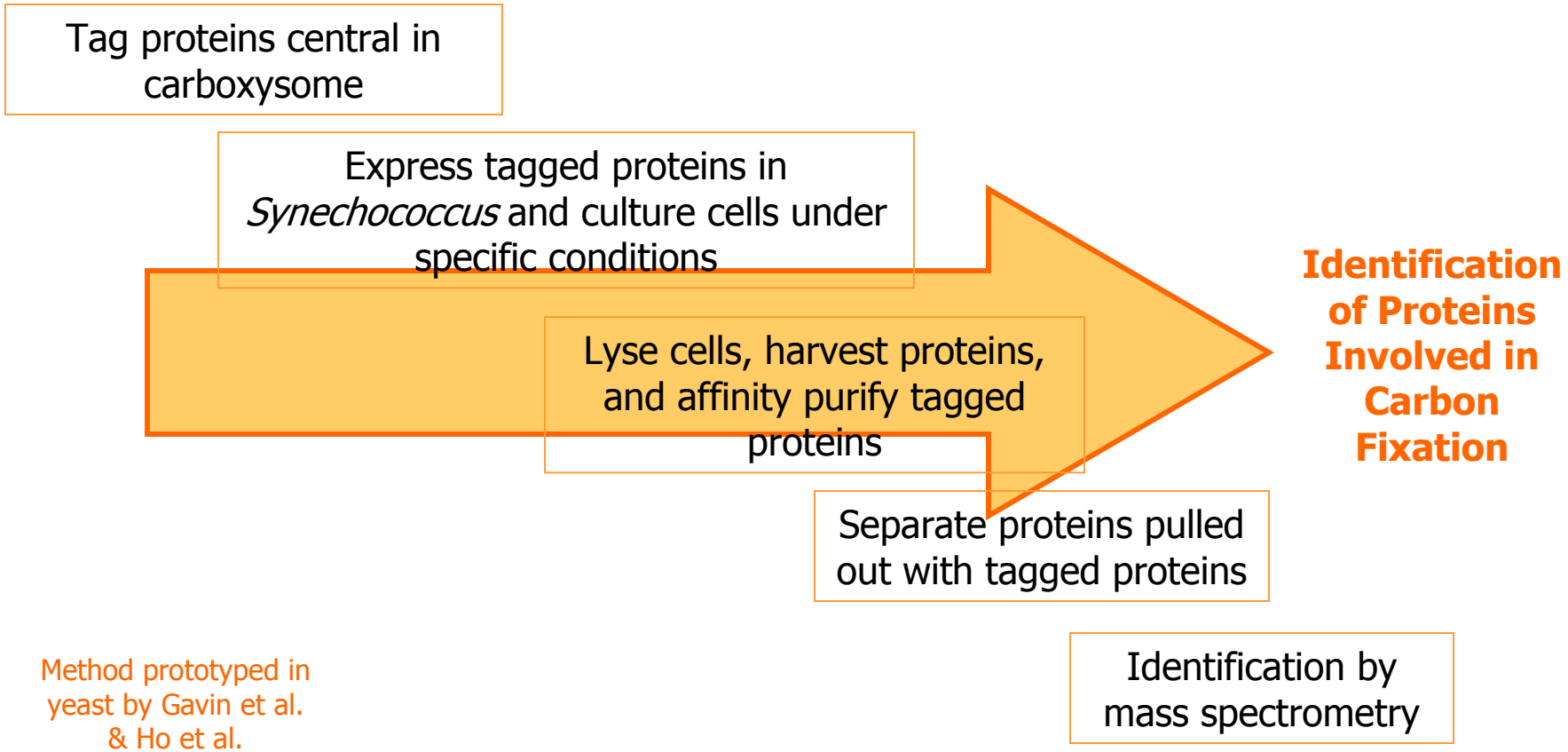


Kaplan et al., Annu. Rev. Plant Physiol. Plant Mol. Biol. (1999)

Characterize Carbon Fixation Complexes

Affinity Purification Mass Spectrometry Strategy

Goal: Spatial and temporal characterization of key carbon fixation complexes as a function of carbon and nutrient concentrations



Method prototyped in yeast by Gavin et al. & Ho et al.

Characterize Protein Binding Domains

Target: *Synechococcus* Carbon Fixation Protein Complexes

Sequence data of binding ligands provides structural information and universally applicable recognition rules used to infer entire protein networks.

Experimental Approach

- Use phage display to determine binding domains in protein complexes
- Analyze binding domains known to exist in prokaryotes to establish binding ligands.
- Use naturally occurring ligands in expression library screening to discover new binding domain proteins.
- Verify binding, binding affinities, and cognate partners with ELISA and Yeast 2-hybrid.

Strategies

Multiple Experimental Techniques

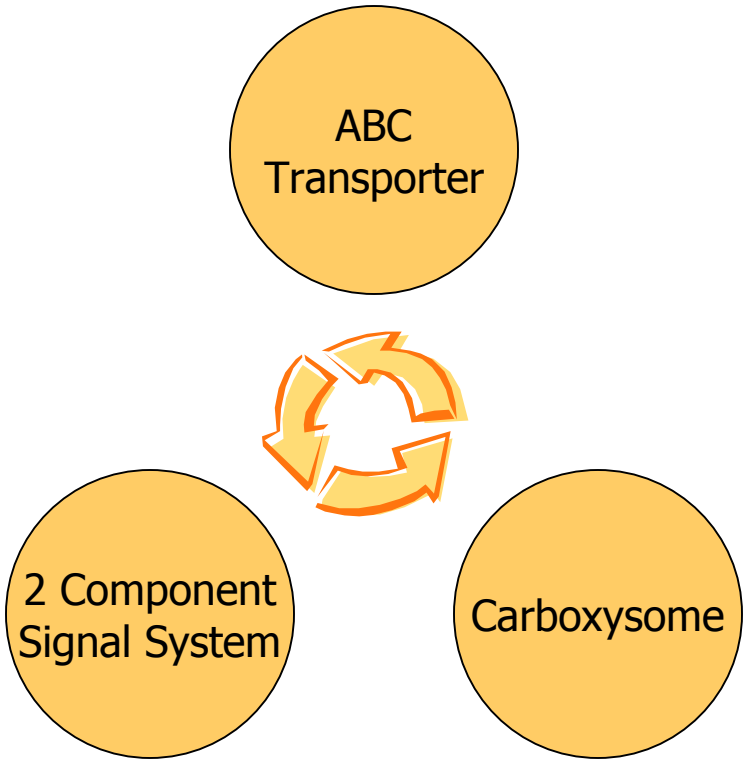
- Phage display
- Yeast 2-hybrid
- Expression library screening
- ELISA

Computational Input

- Molecular biophysics calculations to provide binding affinity rankings for phage display
- Probabilities provided from protein network elucidation

Regulatory Networks

Experimental Strategy



Cis-acting Regulatory Motifs via Microarrays

- 250 gene array
  - ABC transporter proteins
  - 15 kinase-response regulator proteins
  - stress proteins.
- Cluster analysis of induced genes in response to gene knockouts and inorganic substrate levels.
- Identify gene regulatory motifs.

Regulation of ABC transporter Expression

- Make antibodies to 18 substrate binding proteins.
- Test expression to nutrient stresses.
- Develop protein arrays.

Experimental Progress to Date

Carboxysome Analysis

- Optimizing carboxysome preparations through a protocol of high speed centrifugation and sucrose gradation for analysis later by SDS-PAGE and mass spectrometry.
- Protein interactions within the carboxysome are being analyzed by bacterial 2-hybrid, phage display, and affinity-tag pull-down experiments.


SDS-PAGE of protein fractions at various stages of the carboxysome preparation.

Gene Regulatory Analysis

- Microarray experiments to determine gene expression profiles as a function of phosphate limitation and histidine kinase-response regulator knockouts are in progress.
- Full *Synechococcus* WH8102 genome microarray chips are being built.

Individual operon genes PCR amplified and later cloned

Protein-Binding Motif Analysis

- Phage display experiment protocol for protein binding motifs in *Synechococcus* Sp. WH8102 in progress.

Environments & Infrastructure

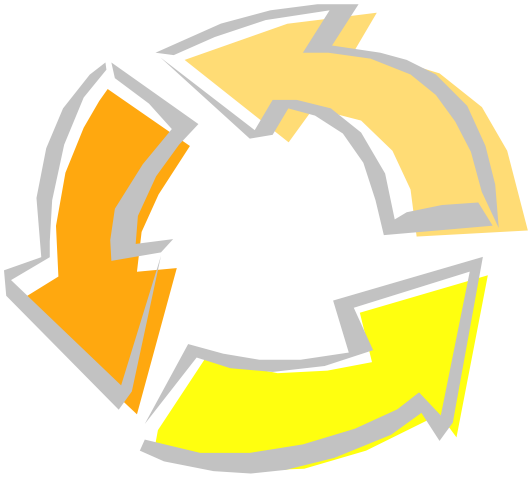
Complements the Experimental & Computational Efforts

Molecular Machines

Regulatory Pathways

Systems Biology

Computational Investigations & Capability Development



Experimental Investigation & Data Generation

Work Environments & Computational Infrastructure

Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Computational Biology Environments & Tools

- Develop working environments with transparent access to distributed databases and computational resources
  - Biology web portals
  - Electronic lab notebooks
- Create new GTL-specific functionality for the work environments
  - Graph data management for biological network data
  - High-performance clustering methods
- Efficient data organization and processing of microarray databases
- High-performance computational infrastructure for biology

Computational Biology Work Environments and Infrastructure

Web-based Tools

- Online Synechococcus data base under development: currently have ORACLE database with more than 180 whole genome annotations. (2 TB on order)
- SVMMER protein functional characterization web portal (http://www.csm.ornl.gov/comp_biology/projects/SVMMER/)
- Prototype Biopathways Graph Data Manager (http://www.lbl.gov/~olken/graphdm/graphdm.htm)
- Pattern analysis tool (PAT) web portal, for statistical comparative analysis of protein-protein interfaces, surface patches and core; protein functional elements (e.g. active/binding sites, DNA-binding sites); and various patterns derived from structural and sequence information. (http://www.csm.ornl.gov/comp_biology/projects/PAT/)

Sandia's GTL Project

For More Information

www.genomes-to-life.org

Participants

Sandia National Laboratories
Bioinformatics & Data Visualization
Experimental Biology
Spectroscopy & Multivariate Analysis
Computational Molecular Biology
Complex Systems Modeling
Statistics & Experiment Design
High Performance Computing

Oak Ridge National Laboratory
Bioinformatics
Computational Molecular Biology
Statistics
High Performance Computing

Lawrence Berkeley National Laboratory
Data Management

Los Alamos National Laboratory
Computational Molecular Biology

National Center for Genome Resources
Complex Systems Modeling

Scripps Inst. of Oceanography, UCSD
Experimental Biology

Joint Institute for Computational Science
Computational Science
High Performance Computing

University of Michigan
Experimental Biology

The Molecular Science Institute
Complex Systems Modeling

University of California, Santa Barbara
Bioinformatics

University of Illinois
Computational Molecular Biology

Sandia National Laboratories

Sandia, Oak Ridge, Lawrence Berkeley, Los Alamos, Molecular Science Institute, Joint Institute for Computational Science
National Center for Genomic Resources, UC San Diego, UC Santa Barbara, Illinois, Michigan

Sandia's GTL Project

Acknowledgements

Grant S. Heffelfinger[1*], Anthony Martino[2], Andrey Gorin[3], Ying Xu[3], Mark D. Rintoul III[1], Al Geist[3], Hashimi M. Al-Hashimi[8], Laurie J. Frink[1], Andrey Gorin[3], William E. Hart[1], Erik Jakobsson[7], Todd Lane[2], Brian Palenik[6], Steven J. Plimpton[1], Diana C. Roe[2], Nagiza F. Samatova[3], Charlie E. M. Strauss[5]

*Author to whom correspondence should be addressed (gsheffe@sandia.gov)

[1]Sandia National Laboratories, Albuquerque, NM
[2]Sandia National Laboratories, Livermore, CA
[3]Oak Ridge National Laboratory, Oak Ridge, TN
[4]Lawrence Berkeley National Laboratory, Berkeley, CA
[5] Los Alamos National Laboratory, Los Alamos, NM
[6] University of California, San Diego
[7] University of Illionois, Urbana/Champaign
[8] University of Michigan

Sandia's GTL Project

Collaborators

- ORNL-PNNL GTL: Michelle Buchanan, Tema Friedman, Jane Razumovskaya, Dong Xu, Gregory Hurst, Robert Hettich, Nathan Verberkmoes
- UC Santa Clara: Carol Rohl
- Scripps Research Institute: Ruben Abagyan
- The Institute for Genomic Research: Ian Paulsen
- Australian National University: Murray Badger and Dean Price
- Cal Tech: Grant Jenson
- ANL: Natalia Maltsev
- The Center for the Advancement of Genomics (TCAG): Marshall Peterson, Scott Collins
- University of New Mexico: Margaret Werner-Washburne